A Convolutional Neural Network Approach for Objective Video Quality Assessment

Patrick Le Callet, Christian Viard-Gaudin, Dominique Barba

A Convolutional Neural Network Approach for Objective Video Quality Assessment

P. Le Callet, C. Viard-Gaudin, and D. Barba

Institut de Recherche en Communications et Cybernétique de Nantes – UMR CNRS 6597

Keywords: convolutional neural network, video quality assessment, MPEG 2, temporal pooling

Corresponding author:

Christian VIARD-GAUDIN

IRCCyN – Ecole Polytechnique de l'Université de Nantes
Rue Christian Pauc – BP 50609
44306 NANTES Cedex 3 – France
Tel : 33.2.40.68.30.40
Fax : 33.2.40.68.32.32
Email : Christian.Viard-Gaudin@univ-nantes.fr

Abstract

This paper describes an application of neural networks in the field of objective measurement method designed to automatically assess the perceived quality of digital videos. This challenging issue aims to emulate human judgment and to replace very complex and time consuming subjective quality assessment. Several metrics have been proposed in literature to tackle this issue. They are based on a general framework that combines different stages, each of them addressing complex problems. The ambition of this paper is not to present a global perfect quality metric but rather to focus on an original way to use neural networks in such a framework in the context of reduced reference quality metric. Especially, we point out the interest of such a tool for combining features and pooling them in order to compute quality scores. The proposed approach solves some problems inherent to objective metrics that should predict subjective quality score obtained using the single stimulus continuous quality evaluation (SSCQE) method. This latter has been adopted by VQEG (Video Quality Expert Group) in its recently finalized RRNR-TV (Reduced Referenced and No Reference) test plan. The originality of such approach compared to previous attempts to use neural networks for quality assessment, relies on the use of a convolutional neural network (CNN) that allows a continuous time scoring of the video. Objective features are extracted on a frame-by-frame basis on both the reference and the distorted sequences, they are derived from a perceptual-based representation and integrated along the temporal axis using a Time Delay Neural Network (TDNN). Experiments conducted on different MPEG-2 videos, with bit rates ranging from 2 to 6 Mbits/s, show the effectiveness of the proposed approach to get a plausible model of temporal pooling from the human vision system (HVS) point of view. More specifically, a linear correlation criteria, between objective and subjective scoring, up to 0.92 has been obtained on a set of typical TV videos.

I Introduction

Video systems, which television programs are an important specific case, are produced for the enjoyment or education of human viewers. Thus, their opinion about the visual quality of such videos is of prime importance. Speaking of quality does not relate here to artistic beauty or sensitive content but just relies on perception of picture distortions from the original scenes as they have been recorded by the scanning camera. Modern video systems are composed of many different stages throughout the production and distribution chain, each of them could be responsible for introducing various kinds of distortions within the video. As a matter of fact, it is often required to convert the video signal into a variety of signal types including non-linear compressed forms. The television signal has to be compressed for storage, efficient transmission, or intra-facility interconnection in digital form. Typically, MPEG compression standard is used resulting in an MPEG transport stream (MTS) which is then multiplexed with other MPEG transport streams for transmission or interconnection in order to optimize the transmission bandwidth. At the receive end of a transmission system, the desired program is demultiplexed from the MTS and the program data is decompressed. With classical coding schemes, it is possible to provide different video picture quality levels based on bit rates. Distribution quality to the home may be adequate using MPEG2 with bit rates from 2 to 5 Mbits/sec for standard definition television (SDTV) and 15 to 19 Mbits/sec for high-definition television (HDTV). However, it is not possible to directly link the perceived quality to the bit rate. Effectively, two different video contents compressed at the same bit rate, will not produce the same perceived quality after decoding. In addition to distortions due to lossy compression algorithms that occur at the distribution network head, trans rating nodes inside the network produce some distortions. In this paper, we restrict the distortions to the coding artifacts produced by a MPEG-2 coding scheme.

Quality assessment is achieved using two types of methods: objective or subjective. The really important point is the opinion of the viewer about the quality of the video, this is why formal subjective tests have been developed for many years [1]. With the advent of digital video compression, the number of different test methods in BT.500 have increased every year. In the past decades, many objective quality metrics for measuring video impairments have been investigated. There is general agreement that there are three methodologies for objective picture quality measurement that provide three levels of measurement accuracy [2]. They are identified as follows:

- Full-reference (FR) metrics do a comparison between a reference video and the tested video; they require the entire reference video to be available, usually in uncompressed form, which is quite an important restriction on the usability of such metrics. Nevertheless, FR metrics should be the most accurate ones since they handle the whole reference sequence. Ideally, FR metrics should be robust regarding the different kinds of distortions in order to benchmark image processing systems.
- Reduced-reference (RR) metrics usually extract a number of features from the reference video (e.g. amount of motion, spatial detail), and the comparison with the tested video is then only based on those features. First intensively studied [3] by Institute for Telecommunication Sciences (ITS), RR metrics are very useful to monitor quality on transmission network, in such context the reduced reference is transmitted with the coded sequence assuming that the reduced reference corresponds to a reasonable overhead. At the receptor side, the coded sequence is decoded in order to compute its reduced representation. The quality is obtained by comparing the reduced representation of both distorted and reference sequences.

- No-reference (NR) metrics exploit only the video under test and have no need of reference information. This allows to measure the quality of any video, anywhere in an existing compression and transmission system.

Figure 1 illustrates these three types of method.

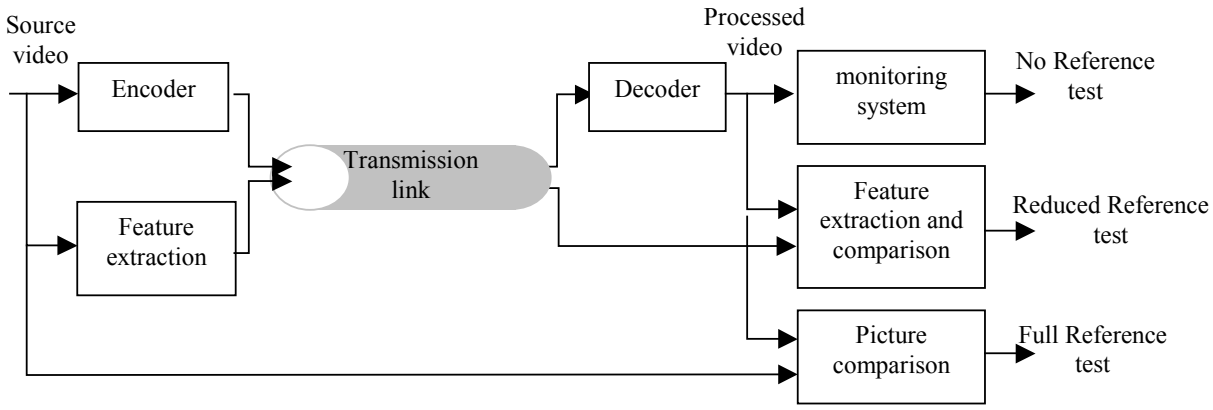


Figure 1. Three types of objective method test

The results presented in this paper concern the field of RR metrics. Classical objective RR metrics of the literature are based on a common framework. We propose here an original method using neural network to tackle some issues in this general framework, especially concerning the last stage. For the needs of the study, we have to define a complete metric but we have voluntarily designed a quite simple front stage (features extraction) to outline the efficiency of the proposed technique for the considered stage. So, the overall proposed quality metric is not optimal. Other quality metrics proposed in the literature based on the same framework but probably more sophisticated in their front stage (features extraction) could take benefit of this technique.

VQEG (Video Quality Expert Group) is the well known main contributor to the normalization of video quality metrics. It has recently finalized a RRNR-TV test plan. This test plan organizes the condition of the competition between objective Reduced Referenced and No Reference quality metrics for TV sequence. To compare metric performances, subjective

quality scores are required, this is why an experimental methodology is necessary. The single stimulus continuous quality evaluation (SSCQE) method has been elected that can lead to some problems for objective metrics. Most of literature metrics are designed to output a single quality estimation for a given video sequence, therefore, they are not supposed to replicate the process of continuous quality estimation as it is performed by human observers. The proposed neural network tool brings some answers to these problems and so can be useful for the future normalization in that field.

The remainder of this article is organized as follows. In section II, we present the joint problematic associated with subjective protocols and objective metrics and we review several works related to objective quality systems. Then, we provide an overview of the proposed system (section III) and explain our reduced video representation in section IV, whereas details about the neural network architecture are given in section V. We describe the databases used to train and test the system in VI. The quality assessment performance of the proposed system is evaluated on a large dataset in section VII. Section VIII concludes the paper with an outlook on future work.

II Video quality assessment

II.1 Subjective methodology and consequences for objective metrics

Advantages of subjective testing are: a test may be designed to accurately represent a specific application; valid results are produced for both conventional and compressed television systems; a scalar mean opinion score (MOS) is obtained; and a wide range of still and motion picture applications are accommodated. In the experiments related later in this paper, we will use the DMOS (Difference Mean Opinion Score), which is the difference of MOS obtained on the reference sequence and on the distorted sequence respectively. A low DMOS means

little degradation whereas an important value corresponds to severe distortions in the sequence.

Weaknesses of subjective testing are: a wide variety of possible methods and test element parameters must be considered; meticulous set-up and control are required; many observers must be selected and screened, and complexity makes it very time consuming. The results of subjective tests are only applicable for development purposes; they do not lend themselves to operational monitoring, production line testing, trouble shooting, or repeatable measurements required for equipment specifications.

The need for an objective testing method of picture quality is clear, subjective testing is too complex and the results provide too much variability. However, since it is the observer's opinion of picture quality that is important, any objective measurement system must be in good correspondence with subjective results for the same video system and test sequences. It means that the goal of an objective metric is to mimic observers behaviour confronted with an experimental protocol. In other words, an objective metric is highly dependent with protocols of subjective quality assessment to prove its efficiency. In the previous VQEG test plan for full reference TV, the Double Stimulus Continuous Quality Scale (DSCQS) method was used for subjective testing. In the DSCQS method, the observer is asked to evaluate the picture quality of sequences using a continuous grading scale and to give one score for each sequence. This is maybe the reason why most of objective metrics in literature are designed to output a single quality score for a given video sequence (typically 8 seconds long). Though they can generate frame by frame scores, they are not suited to replicate the process of continuous quality estimation as it is performed by human observer with the SSCQE method. In this latter case, a digital video sequence (usually several minutes long) is presented once to the subjective assessment viewer (the video sequences may or may not contain impairments). Observers evaluate the picture quality in real time using a slider device (typical sampling rate

of 2 per second) with a continuous grading scale composed of the adjectives Excellent, Good, Fair, Poor and Bad. This methodology has been chosen by VQEG for RR-NR TV test plan essentially because it is consistent with real-time video broadcasting where a reference sample with no degradation is not explicitly available to the viewer. Nevertheless, it induces some observer's behaviours that can be very challenging to mimic for objective metric. Two main effects are identified:

- Response time delay: human observers make decision and displace the slider to reflect their opinion. The consequence is a delay between the moment the displayed frames and the corresponding right position of the slider. Ideally, the objective and the subjective results should be synchronized. Unfortunately, the delay is not constant, it depends on many factors. We suspect that delay's variation is mainly due to the content and the temporal variation of the distortions.
- Asymmetric tracking: in general humans experience greater feelings intensity from disliked situations compared to favorable ones. In other words, observers criticize quickly and forgive slowly. This leads to an asymmetric tracking of subjective score with the SSCQE metric : MOS takes les time to fall when distortions appear than to raise when distortions disappear.

II.2 Objective quality metrics

Usually, FR or RR metrics are composed of two main stages. In the first one, the errors between original and distorted images are computed. In the FR metrics case, it leads to distortion maps whereas, in the of RR metrics case, it deals with the difference between features that constitute the reduced representations. The second main function allows to pool the errors or the differences, and thus, to provide the global quality assessment. This second function is highly dependent on the subjective protocols. As a matter of fact, a good metric

should be well balanced between error visibility stage and error pooling. We have previously demonstrated [4] the complementarity of these two stages.

Two categories of image quality metrics can be found in literature. Metrics from the first category try to exploit the properties of known artifacts, such as blocking artifacts, using feature extraction and model parameterization [5]. This class of metric focuses on the particular type of artifacts [6]-[8], so it is not universal. In someway, these specialized metrics can tackle some problems inherent to distortion weighting, but they do not bring a complete answer for error pooling regarding subjective protocols. For NR metric, the task is even harder therefore few works are present in literature. In [9] for example, authors propose a NR metric for compressed picture (DCT and block based scheme) to reach PSNR performances. Metrics from the second category, such as proposed in [10], [11] and [12], use a human visual system (HVS) model for low level perception, such as sub-band decomposition and masking effect, in order to compute distortion maps. Most of these approaches use psycho-visual models stemming from psychophysics experiments. Recently, we have explored such approaches for RR metric providing a way to produce reduced representation according to low level perception mechanisms [13]. The main limitation of a HVS based metric is due to the lack of knowledge to model the error pooling process. Effectively, since it is difficult to address high level perception mechanisms through experiments and as the pooling stage is connected to these mechanisms, these metrics suffer from the lack of data to be coherent all along their processing steps.

For FR metric, some interesting ideas, even if they are not linked to psychophysics experiments, have been proposed as an alternative to the conventional but not realistic Minkowski summation. In [14] a structural approach is used in order to predict DSCQS subjective score. In [15], an original cognitive emulator, based on rational analysis, provides a simulation of high level processing of visual information in the context of SSCQE method.

The method has been evaluated using three sequences coded at three rates (MPEG2 MP@ML) leading to a total of 9 minutes. Face to the problem of the variable response time, authors claim that they cannot use reliable metric to compare their metric output with SSCQE results, therefore they simply present graphs. They argue that theirs results are better than PSNR. In [16], a temporal summation stage based on a recursive formulation is used to combine distortion across frame in a way that effectively models recordings from human observers with SSCQE method. It is a low pass FIR filter and it takes into account the fact that viewers do not respond equally to increasing and decreasing changes in the perceived distortion. In order to evaluate the metric, eight reference sequences, 30 seconds long each, have been coded at two different bit rates using three coders in order to generate sequences to be assessed by observers. Subjective materials have been split in order to provide several training sets of video frame to tune some metric parameters. Comparisons between the metric's output time series and the SSCQE recordings are done with a specific fitted Mean Square Error (MSE). Authors claim that distance measures such as usual MSE are too much affected by offset between the two score series, and that measures such as correlation can be affected by small variations over time, in spite of the overall similarity of the two time series. The proposed metric generally performs roughly as well, or even better, as a MSE based metric.

For RR metric, a neural network approach [17] has been proposed in order to mimic human pooling in the context of DSCQS subjective protocol. This system process a 20-input feature vector that is forwarded to a radial basis function neural network (RBFNN) for classification. A NR metric is based on a CBP neural networks to pool feature in [18] for SSCQE protocol. The idea is very promising even if the performance of the metric has not been assessed with usual measure. To the best of our knowledge, an original pooling method corresponding to SSCQE protocol for a RR metric has not been yet proposed.

III Overview of the proposed system

Most of our previous metrics were HVS-based [19][20][21], but we focus in this study on the pooling process adapted to SSCQE for a RR metric. The only HVS property considered here comes from feature extraction, which is carried out on a perceptual color representation of the video sequence. Color can be very useful in quality assessment. In order to limit redundancy between components, it can be interesting to choose carefully the color space as it introduces negligible computing complexity. Krauskopf's color space [22] has been selected since we have previously validated it through psychophysics experiments conducted in our lab [23]. Therefore, YUV original images are transformed into three perceptual components: A (Achromatic), Cr1 (red-green axis) and Cr2 (yellow-blue axis).

The design of a RR quality assessment system needs to define two main sub-systems: i) construction of the information that has to be extracted both from the reference video and the decoded video, ii) comparison of the two feature sets and pooling.

Perceived quality of video sequences is affected by distortions that are present in the spatial domain, and also by the temporal duration and evolution of these distortions. Although these two contributions are highly interdependent, we will assume a model that first extracts a description vector on a frame by frame basis. That means that the extracted features are spatially integrated, and then, we will consider the pooling of the different features of the vector along the temporal dimension. One can imagine many different features. In the general framework of objective metrics, features choice is as crucial as pooling definition. Since it is not the main goal of this work, we have selected a set of 4 features from the literature. They are well suited in order to sum up the content of a frame. These features are described in more details in section IV, three of them are totally content dependent (regarding frequency and temporal content). The last feature is more focused on distortion a priori related to blocking effect. Each of these 4 features is computed independently on the three perceptual

components. Consequently, the global size of the feature vector describing every frame is $3 \times 4 = 12$ features.

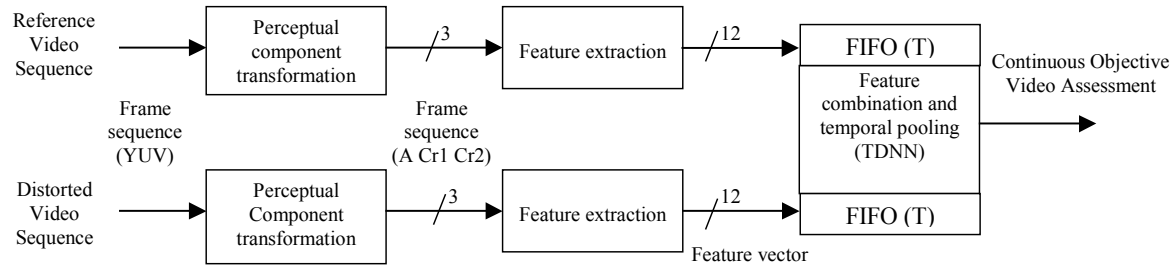


Figure 2. General scheme for the RR objective video quality assessment

The last stage of the system, presented in Figure 2, is the main contribution of this study. It corresponds to the feature combination and the temporal pooling of the feature vector sequence. As we have mentioned before, the design of this stage is not straightforward. To overcome these difficulties, we propose to base this function on a learning algorithm that will be able to generalize the observed behavior from a collection of subjective tests. We introduced a neural net (NN) approach using a constrained architecture that is well suited to mimic not only the temporal integration of distortions but also to construct new measures of distortions from initial features. This explains why we have selected content based features rather than distortions based features. As explained later, the distortions will be constructed by the first layers of the TDNN, combining content features from the distorted and the reference sequences.

The TDNN architecture is more precisely detailed in section V. It corresponds to a time delay neural network (TDNN), which performs convolution functions on the video sequence. It allows to model the following behaviors: 1) systematic local analysis to construct meta distortions 2) assessor's reaction times are subject to delays; 3) time-consecutive frames tend

to interfere with one another, and 4) the most recent frames of a sequence have a greater effect on the overall quality rating.

Two main temporal parameters have to be defined when scanning the video sequence, see Figure 3. The first one, Δ, is related to the refreshing rate of the quality assessment. For this study, the rate is two subjective scores per second to be consistent with VQEG RR-NR TV test plan. The second parameter, T, takes into account the number of previous frames that will affect the perceived quality at time t, resulting in a grading G_t. This is an important point of the proposed method. Not only the present frame will participate to the computation of the objective grading G_t, but a memory function has to be integrated to mimic the behavior of the human visual system that is sensitive to the sequential nature of the video sequence [24].

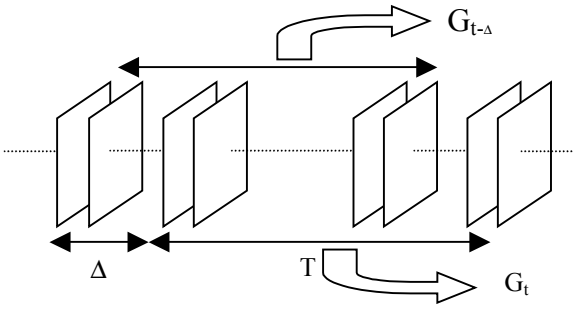


Figure 3. Video grade updating

IV Features extracted from the frames

We propose to select the features directly from existing objective metrics mainly proposed for the FR-TV VQEG phase I. Even though these metrics exhibit poor correlation with human judgment when standard pooling stages are used, we would like to use the same kind of features and experiment what the proposed NN approach can do in this context. As explained in the previous section, we are not interested by the capability of these features to be used as explicit distortion features. Distortion measure will be achieved in the early layers of the TDNN through comparison of content based features between original and distorted

sequence. Therefore, some adaptations are performed from literature features in order to get rid of the explicit comparison process. Based on this principle, the four features are described in the following sub-sections.

IV.1 Frequency content measures: GHV and GHVP

The two first features, termed as *GHV* and *GHVP*, are derived from the work presented in [25]. They have been elaborated to detect the blurring artifacts but are also sensitive to tiling distortions. These two features are computed from the two-dimensional histogram *SIH(r, θ)* where *r* is the magnitude of the gradient vector, and θ is the orientation of the gradient vector with respect to the horizontal axis and *SIH(r,θ)* is the number of pixels in the gradient image whose gradient radius and angle is *r* and θ, respectively.

The feature *GHV* whose value increases as the number or sharpness of horizontal and vertical edges increase is given as:

$$GHV = \frac{1}{p}\sum_{r}\sum_{\theta} SIH(r,\theta).r \quad \text{with: } 0 < c_a \le r \le c_b \text{ and } \theta = \frac{k\pi}{2},\ (k = 0,1,2,3) \qquad \textbf{Equ. 1}$$

where *r* and θ are as defined above and c_a and c_b are clipping limits, and *p* is the number of pixels in the image.

In order to separate blurring from tiling, the *GHVP* feature that characterizes the edge content of the image *without* the inclusion of horizontal and vertical edges is also computed:

$$GHVP = \frac{1}{p}\sum_{r}\sum_{\theta} SIH(r,\theta).r \quad \text{with: } 0 < c_a \le r \le c_b \text{ and } \theta \neq \frac{k\pi}{2},\ (k = 0,1,2,3) \qquad \textbf{Equ. 2}$$

IV.2 Temporal content measure: Power of frame difference

The next extracted feature, P, is based on temporal changes in sequence. First introduced in [26], such type of information is very useful in video quality assessment. It has also been exploited in [27]. In this latter purpose, authors consider the following distortions: flicker,

jadder, moving blurred images, random noise and edge jitter, and define linear combinations of some distortion factors using properties of visual perception. These combinations, which are explicitly defined in their work, are based on the power of the frame difference images computed respectively on the original and on the distorted video sequences. In our work, we will just keep the computation of the power of the frame difference and use it as an input feature for the NN. It will be the responsibility of the NN to model the distortions. The following computations are proceeded:

Frame difference: $$d(t, m, n) = I(t, m, n) - I(t-1, m, n)$$ **Equ. 3**

Power of frame difference: $$P = \sum_{m,n}^{all} \left\{ d\left(t,m,n\right) \right\}^2$$ **Equ. 4**

IV.3 Blocking measure: B

This last measurement is mainly dedicated to exhibit blocking effects [28]. It is based on the method described in Figure 4 and proposed in [29].

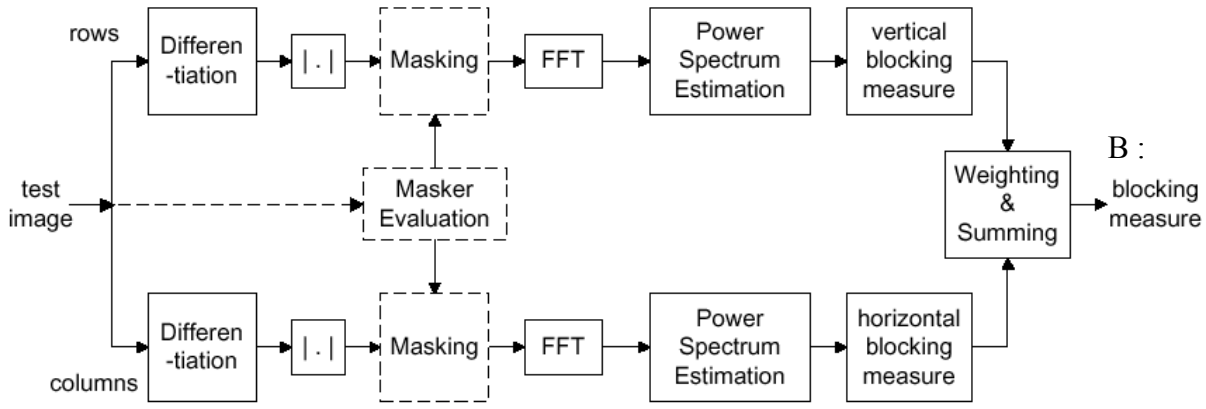


Figure 4. Computation of B feature

They apply 1-D FFTs to horizontal and vertical difference signals or rows and columns in the image to estimate the average horizontal and vertical power spectra. Peaks in these spectra due to 8×8 block structures are identified by their locations in the spectra. The power spectra of the underlying non-blocky images are approximated by median-filtering these curves. The

overall blockiness measure, feature B, is then computed as the difference between these power spectra at the locations of the peaks. Integration of masking effects is possible with this scheme while it has not been used in our implementation.

Accordingly, for every new frame we compute the four features (*GHV, GHVP, P, B*) previously described on each of the three perceptual components A, Cr1 and Cr2. Hence the input vector for the NN has a size of $(4 \times 3 \times T) \times 2$ for a RR system, where T is the number of frames taking into account in the computation of a scoring, as displayed in Figure 2.

V Neural network architecture

The ability of multi-layer networks with gradient descent to learn complex, high-dimensional, non-linear mappings from collections of examples makes them obvious candidates for many tasks related to machine vision systems. A recent survey [30] denotes more than 200 applications of NN to image processing. They can address most of the various steps which are involved in the processing chain: from the preprocessing/filtering to the image understanding level.

Multi Layer Perceptron (MLP) are the most common neural network architecture encountered. They consist of several layers of fully-connected hidden units. However, when the number of input variables is quite large, which is the case with image application, this architecture leads to several tens of thousands of weights. Such a large number of parameters increases the capacity of the system but at the same time requires a larger training set. In addition, the memory requirement to store so many weights may rule out certain low capacity system such as mobile device. To overcome the dilemma between small NN with low capacity and large NN that appear overparameterized with respect to the size of the training database, one can design specific architectures that aim to detect and combine local features.

The idea is to perform the same kind of computation at every place in the video stream based on a local receptive field. This is typically the principles involved with convolutional NN (CNN). Introduced by LeCun et al [31] and successfully used in different domains [32], they are powerful bioinspired hierarchical multilayered neural networks that combine three architectural ideas: local receptive field, shared weights, and spatial or time subsampling.

In our case, the convolution kernels will be defined along the temporal axis, leading to the so-called Time Delay Neural Network (TDNN). TDNNs, which were previously applied to speech recognition [33] and handwriting character recognition [34], are well suited to sequential signal processing. They allow to preserve the sequential nature of data, in contrast with standard MLP where the topology of the input is entirely ignored. On the contrary, video sequences have a strong local structure: frames that are temporally nearby are highly correlated. Local correlations are the reasons for the well-known advantages of extracting and combining local features before processing temporal objects. With CNN, a given neuron detects a particular local feature of the video stream. It performs a weighted sum of its inputs followed by a non-linear squashing function (sigmoid). Its receptive field is restricted to a limited time window. The same neuron is reused along the time axis to detect the presence or absence of the same feature at different position of the video stream. A complete convolutional layer is composed of several feature maps, so that multiple features can be extracted at each temporal position. This weight sharing technique greatly reduces the number of free parameters and hence trained networks run much faster and require much less memory than fully connected NN.

The idea of connecting units to local receptive fields on the input was largely inspired by Hubel and Wiesel's discovery [35] of locally-sensitive, orientation selective neurons in the cat visual system and local connections have been used many times in neural models of visual learning, [36], [37]. With local receptive fields, neurons can extract elementary visual

distortions in videos. These distortions are then combined by the subsequent layers in order to detect high-order features.

In addition to the TDNN layers, the upper layers are standard fully connected layers. With this application, the last layer consists of a single neuron fully connected to the previous layer; the output of this neuron will be trained to estimate the DMOS value as it has been provided by human observers.

A detailed view of the TDNN architecture is presented in Figure 5.

From this general architecture, many parameters have to be defined to customize a specific learning machine. The most important ones are :

- Local feature extraction sub-system (TDNN type):
    - nb_layer_tdnn: number of layers of the extraction sub-system,
    - T: size of one layer with respect to the time axis,
    - nb_feat : size of one layer with respect to the feature axis,
    - field: size of the convolution field with respect to the time axis,
    - delay: temporal delay between two convolution fields,
- Global estimator sub-system (MLP type):
    - nb_layer_mlp: number of layers of the fully connected sub-system,
    - nb_neurons: numbers of neurons of the hidden layer.

Different values for these parameters have been experimented and are presented in the result section. However, some of these parameters have been set once for all. For example, the number of layers has been set globally to 4, including 2 layers for the local feature extraction sub-system, and 3 for the fully connected NN at the upper level, which correspond to one input layer – actually, the output layer of the TDNN sub-system, one hidden layer and an output layer with a single neuron. The value of T, which refers to the number of frames involved in the computation of a score, has also been kept to the same value (except in Table

6); we supposed that at last the 5 last seconds influence the perceived quality at a given time, consequently, we set T to 5s × 25 f/s = 125 frames.

One important practical issue with these trainable systems is the requirement of a large database. Up to now, only small sets of images with subjective quality marks were available, they did not allow to learn the large number of parameters involved in a NN-based system. Data presented in section VI, appears to fill in the gap and make neural network approaches, and more specifically TDNN, attractive to propose a solution to video quality assessment.

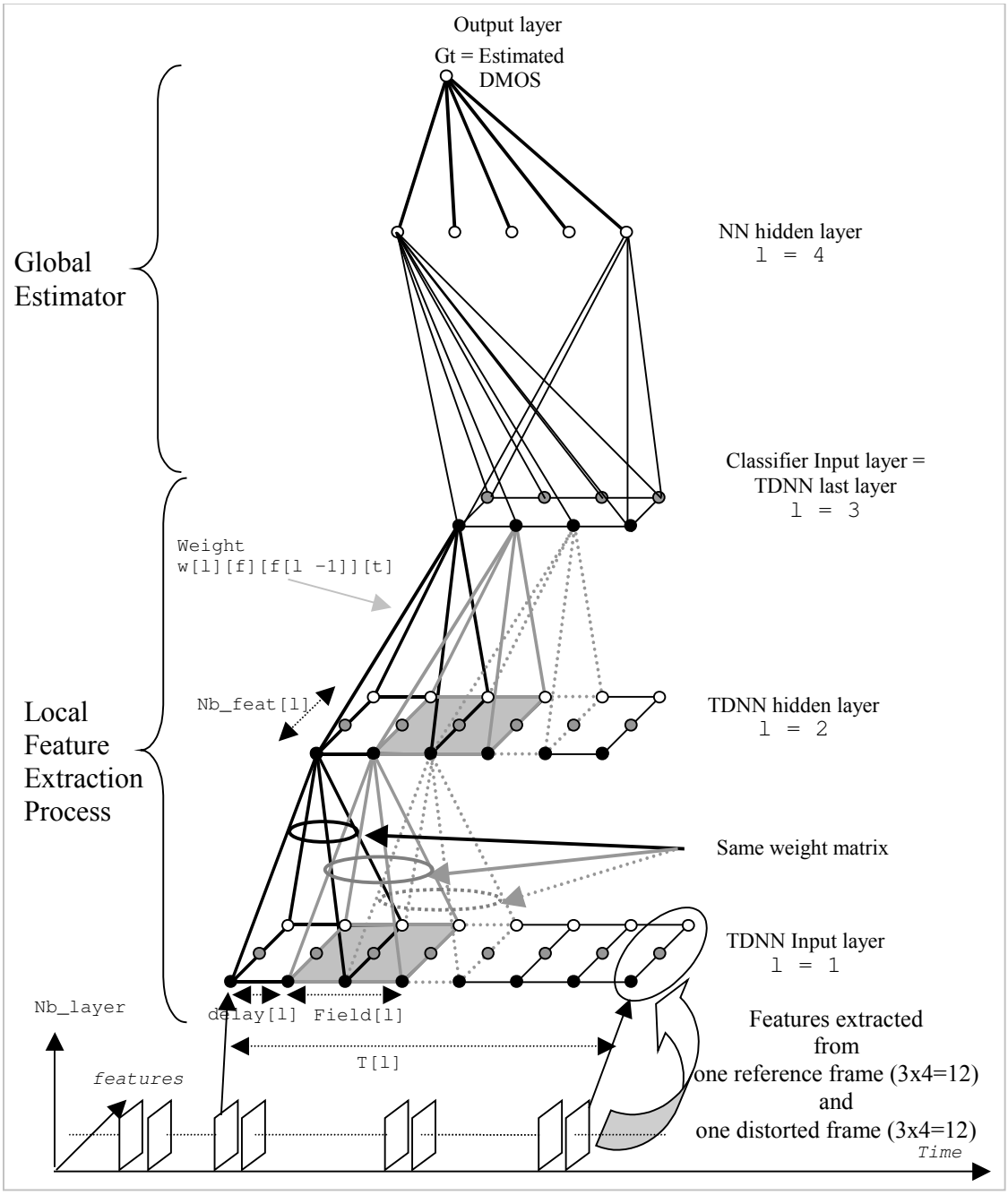


Figure 5. Generic TDNN architecture

VI Material available for training and testing

The database used to train and test the system described in this paper was constructed from materials delivered by TDF[1]. A first set of 4 video sequences (*Cooking*, *Football*, *Horses*, and *Road*) will be considered as 4 reference sequences, each one being composed of 4,500 images

[1] Télédiffusion De France, Research Center Metz

(720 x 576 pixels) that represent a 3 minutes video at a frame rate of 25 images per second. *Cooking* video presents a famous French woman Chief at work, it is an indoor video sequence. The three others are outdoor videos with quite different contents. They represent almost uniform content only from a very high semantic point of view. During the 3 minutes, sequences are not homogeneous in terms of spatial and temporal content comparing with 8 seconds sequences usually used in the DSCQS protocols. They have been produced by a MPEG-2 codec at a bit rate of 8 Mbits/s which ensures a very high quality, very close to the original videos.

Table 1. Distorted video sequences

<table>
<tr><td>Sequence name</td><td>Cooking</td><td>Football</td><td>Horses</td><td>Road</td><td># videos/ subj. rating values</td></tr>
<tr><td>Bit rates in Mbits/s</td><td>2/3/3.5/5</td><td>2/3/3.5/4/5/6</td><td>2/2t/3/3t</td><td>2/3/6</td><td>17/ 17x180x2= 6120</td></tr>
<tr><td rowspan="2">Number of videos/subjective rating values</td><td>4/ 4 x 180 x 2 = 1440</td><td>6/ 6 x 180 x 2 = 2160</td><td>4/ 4 x 180 x 2 = 1440</td><td rowspan="2">3/ 3 x 180 x 2 = 1080</td><td rowspan="2">17 videos/ 6120 ratings</td></tr>
<tr><td colspan="3">14/ 14 x 180 x 2 = 5040</td></tr>
<tr><td></td><td colspan="3">Loo training</td><td>Loo test</td><td></td></tr>
</table>

A second set of 17 distorted video sequences has been produced. They correspond to different bit rates, ranging from 2 Mbits to 6 Mbits per second (Table 1). When the same video is available twice at the same video rate (e.g. Horses at 2 Mbits/s), one of these videos is directly obtained by the coding scheme while the other one (2t) has been derived with a transcoding scheme from the reference 8 Mbits video.

For all of these video sequences, TDF have provided the corresponding subjective assessment results obtained with human observers. Subjective tests were running with more than 20 observers using a SSCQE protocol with hidden reference removal in normalized conditions and environment according to recommendations ITU-R BT.500-10. Subjective

scores (MOS) consist of a quality rating sampled twice a second. It is easy to derive DMOS (difference of MOS between two conditions) with an associated Interval of Confidence (IC) obtained according to subjective measurement procedures.

This distorted video database has been split into two subsets: one for training and the other one for testing the generalization performance of the trained system. Furthermore, we have used a leave-one-out (*Loo*) protocol in order to take advantage of all the material available. In such a way, the training set was composed of the videos from 3 out the 4 groups of videos, for example: *Cooking*, *Football*, and *Horses*, (14 videos for a total of 5,040 subjective quality grades as displayed in Table 1). The test set was composed of the remaining group of videos, in this case: *Road* (3 videos for a total of 1,080 objective video grades to compute and compare with the corresponding subjective grades). Then, we shift to another subset of 3 groups for training, and once again after, in order that every group of video has been used for testing. With this procedure, we made sure that the sets of images of the training set and test set come from disjoint video sequences with quite different video contents.

VII Quality assessment results

VII.1 Baseline results with the complete metric

The TDNN training uses a standard stochastic gradient backpropragation algorithm adapted to respect the constraints of weight sharing [38]. The main change here is the computation of the local gradient of the backpropagated error signal with respect to the shared weights. Considering that every feature contains in fact a single neuron with multiple instances, the local gradient for this neuron is simply the summation of the local gradients over all instances of it [31].

The network cost function is expressed as

$$J_t = \left(DMOS_t - G_t\right)^2 \qquad \textbf{Equ. 5}$$

where $DMOS_t$ is the actual subjective score derived experimentally from the panel observers and G_t is the output of the TDNN.

As a measure of performance of the proposed objective scoring method, three main indicators will be presented. One will be the root mean squared error on the test set, defined as

$$J_{rmse} = \sqrt{\frac{1}{N}\sum_{t=1}^{N} J_t} \qquad \textbf{Equ. 6}$$

where N is the number of scores computed on the test video sequences.

The second one being the Linear Correlation Criteria (*LCC*), which expresses the monotony between *DMOS* and objective scoring, it is expressed as

$$LCC = \frac{\sum_{t=1}^{N}\left(DMOS_t - \sum_{t=1}^{N}\frac{DMOS_t}{N}\right)\left(G_t - \sum_{t=1}^{N}\frac{G_t}{N}\right)}{\sqrt{\sum_{t=1}^{N}\left(DMOS_t - \sum_{t=1}^{N}\frac{DMOS_t}{N}\right)^2 \sum_{t=1}^{N}\left(G_t - \sum_{t=1}^{N}\frac{G_t}{N}\right)^2}} \qquad \textbf{Equ. 7}$$

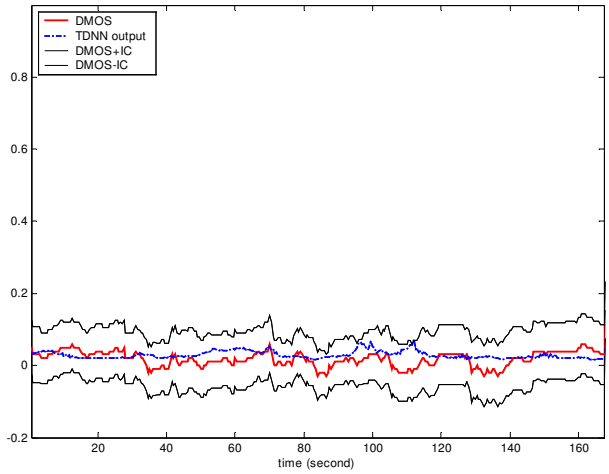
The last measurement will be the percentage of outlier (*OR*), which represents the ratio of objective marks that are outside an interval representing twice the Interval of Confidence (*IC*) from the subjective marks.

The typical behavior of the system on the four different test sets of the *LOO* database, once trained with the complementary training sets of the database as presented in Table 1, is displayed in Table 2.

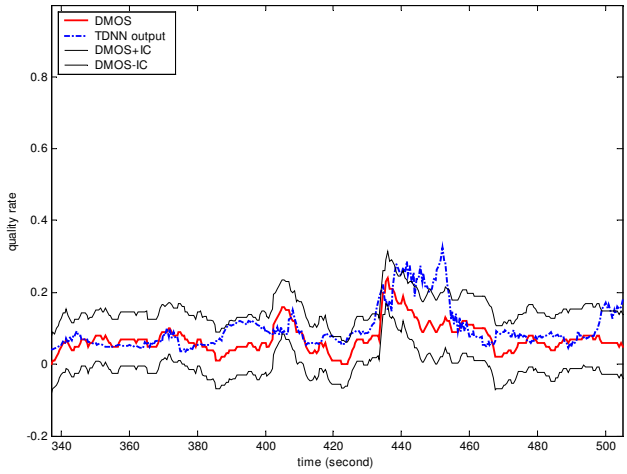
Table 2. Results on the *LOO* Database

Test video content	Number of video scoring	Root mean squared error: *Jrmse*	Linear Correlation: *LCC*	Outlier Ratio *OR* (%)
Cooking	1440	0.086	0.93	3.3
Football	2160	0.092	0.95	9.3
Horses	1440	0.081	0.94	5.6
Road	1080	0.067	0.93	1.3
Global	**6120**	**0.084**	**0.92**	**5.6**

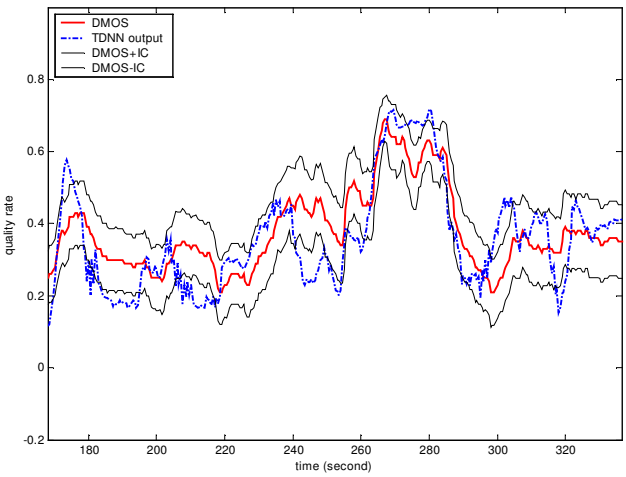
The global set represents a 51-minute video length, which is a very significant amount of time to evaluate the performances of the quality assessment system. Globally, the mean error is less than 10% (8.4%) and the correlation between subjective and objective marks is quite high, it reaches 0.92 on the whole set, and ranges from 0.93 to 0.95 on the individual test sets. The outlier ratio, according to the test video used, varies from about 1% to 10%, with an average value around 5%.



a) *Road* (6 Mbits/s)
Jrmse = 0.0436, *OR* = 0%



b) *Road* (3 Mbits/s)
Jrmse = 0.0558, *OR* = 0.9%



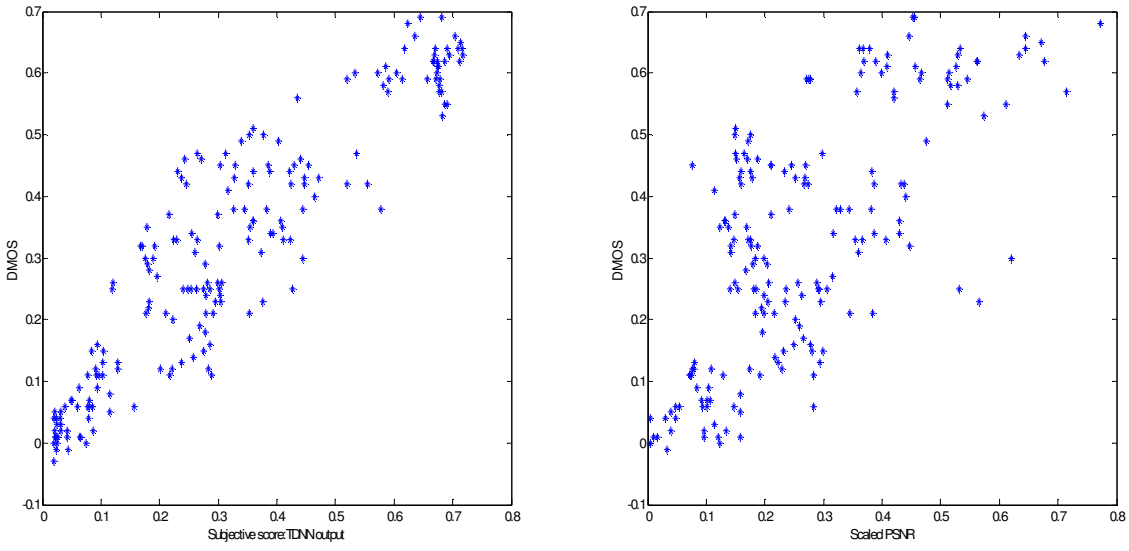
c) *Road* (2 Mbits/s)
Jrmse = 0.0975 *OR* = 3.0%

Figure 6. DMOS (Subjective scoring) and TDNN-based RR System (Objective scoring) on *Road* test set

Detailed results concerning *Road* videos (before next to last row of Table 2) are presented in Figure 6 and Figure 7. In Figure 6, the *DMOS* values are plotted with a continuous line, while

the predicted quality, which is the output of the TDNN, is represented with a dashed-dotted line. Two additional curves are present, they define the incertitude measurement related to the *DMOS* values. They have been set on this chart with a margin equal to the Interval of Confidence (*IC*), which has been computed from the standard deviation of the *DMOS* values taking into account the number of human observers. For every sequence, the value of the root mean squared error between *DMOS* and the output of the TDNN is given (*Jrmse*) with the percentage of the marks given by the TDNN corresponding to outliers ratio (*OR*). Even with the most distorted sequence, *Road* (2 Mbits/s), see Figure 6-c, the predicted quality given by the TDDN-based RR system appears quite satisfying, since still *97%* of the time the predicted output remains inside the margins. This is quite relevant since the subjective score is very variable along this sequence, ranging from *20%* to *70%* of the full scale of distortion, with a very peaky aspect,

On this *Road* test set, the global mean quadratic error *Jrmse* is equal to *0.067* and the correlation criteria *LCC* reaches *0.929*.



$LCC = 0.929$ $LCC = 0.742$

Figure 7. Scatter gram plots: DMOS versus TDNN output and PSNR on the *Road* test set (sub-sampled).

On its left side, Figure 7 displays the corresponding dispersion of the points (Objective marks, Subjective marks). For the sake of comparison, we present on the right side of Figure 7, the dispersion of the points (Scaled *PSNR*, Subjective marks).

While *Peak-Signal-to-Noise-Ratio* (*PSNR*) is a very poor indicator to assess the quality of reconstructed images, it has the advantage of being an easy and well known measurement to evaluate the performance of a compression technique. It is directly derived from the mean square error (*MSE*) computed between a reference image *I(m, n)* and a distorted image *Î(m,n)*:

$$MSE = \frac{1}{MN}\sum_{m=1}^{M}\sum_{n=1}^{N}\left[I(m,n)-\hat{I}(m,n)\right]^2 \qquad \textbf{Equ. 8}$$

$$PSNR = 20\log_{10}\left(\frac{2^p-1}{MSE^{1/2}}\right) \qquad \textbf{Equ. 9}$$

for an image I and a reconstructed image $\hat{I}$, with pixel indices $1 \le m \le M$ and $1 \le n \le N$, image size $N \times M$ pixels, and p bits per pixel.

Charts presented in Figure 7 display, for a sub-set of the *Road* test set, on the *x-axis* the output of the TDNN on the left chart and the scaled *PSNR* on the right chart, and on the *y-axis* the corresponding subjective scores (*DMOS*). In order not to be biased by multiple instances of about the same event, as the sampling rate is quite high (2 samples per second), we have sub-sampled the subjective sequence in order to obtain a quasi-uniform distribution of the marks over the range of *DMOS*.

Points of the left chart have a linear correlation of *0.929* while the right chart has a linear correlation of *0.742*. This value of *0.742* is far below most of the results reported in Tables 2 to 4, hence, the proposed *Reduced Reference* objective video quality assessment method clearly outperforms the *Full reference PSNR* metric.

The TDNN used in this experiment corresponds to the configuration described in the last line of Tables 5 and 6. More specific configurations are studied in section VII.2.

VII.2 Sensitivity analysis of the reduced reference system

a) Sensitivity to the feature set

In section IV, we have introduced a set of 4 features from the literature, termed as GHV, GVHP, P and B. To evaluate the strength and the complementarities of these features, we have conducted experiments where we used individually only one of these features, which is computed on each of the three perceptual components, and for reference and distorted videos. In such a case, the input layer of the TDNN, at a given time, encompasses only 6 inputs instead of 24, corresponding to one feature computed on each of the three perceptual channels for the original and the distorted frames.

From Table 3, it can be observed that each of the features does not perform equally. Feature B in the context of this dataset leads to poor generalization results, it is the worst feature whatever the test video used. Conversely, feature P, related to power of frame difference, clearly outperforms the others on the global set, it allows to achieve the smallest predicted error (0.096) and the highest correlation criteria (0.89) with the subjective scores. However, when considering individually the video content, it is not always the most efficient feature, since on the *Football* video, a slightly better result is obtained using *GHV* feature.

Table 3. Sensitivity to the features, results on the *LOO* test set

Test set	Features	*Jrmse*	*LCC*	*OR %*
Cooking	GHV	0.097	0.88	4.2
	GHVP	0.107	0.81	9.5
	P	0.073	0.95	1.3
	B	0.118	0.77	13.3
Football	GHV	0.126	0.87	16.0
	GHVP	0.126	0.86	17.7
	P	0.128	0.82	21.2
	B	0.137	0.80	19.6
Horses	GHV	0.116	0.83	10.3
	GHVP	0.094	0.89	5.6
	P	0.066	0.95	4.9
	B	0.151	0.73	25.3

Road	GHV	0.136	0.79	27.7
	GHVP	0.110	0.79	13.0
	P	0.080	0.90	3.2
	B	0.147	0.54	46.5
Global	**GHV**	**0.119**	**0.82**	**28.4**
	GHVP	**0.112**	**0.84**	**21.5**
	P	**0.096**	**0.89**	**18.6**
	B	**0.139**	**0.74**	**34.4**

When comparing Table 2 and Table 3, we can notice that the combination of the set of four features boosts the performances obtained with the best single feature; the mean error drops down to 0.084 (instead of 0.096) and the correlation criteria reaches 0.92 (instead of 0.89).

b) Sensitivity to the perceptual components

In this case, we aim to study the impact of the three perceptual components that model the human visual system with respect to their contribution in the perceived quality when using the proposed metric. To achieve these experiments, we keep only one subset of features coming from one perceptual component, as described in Table 4. In such a case, the input layer of the TDNN, at a given time, encompasses only 8 inputs instead of 24, corresponding to four features computed on the corresponding perceptual component of the original and of the distorted frames.

Table 4. Sensitivity to the perceptual components, results on the LOO test set

Test set	Perceptual components	*Jrmse*	*LCC*	*OR %*
Cooking	A	0.058	0.95	3.4
	Cr1	0.061	0.95	5.0
	Cr2	0.062	0.95	4.2
Football	A	0.104	0.95	28.8
	Cr1	0.115	0.92	28.8
	Cr2	0.116	0.94	29.4
Horses	A	0.103	0.88	22.9
	Cr1	0.078	0.93	12.7
	Cr2	0.088	0.96	17.9
Road	A	0.082	0.90	12.3
	Cr1	0.094	0.92	20.3
	Cr2	0.077	0.92	13.4

Global	**A**	**0.091**	**0.90**	**18.5**
	Cr1	**0.092**	**0.90**	**17.9**
	Cr2	**0.092**	**0.89**	**17.9**

We can note that the three perceptual components plays a comparable role in their ability to sum up the visual quality of videos. Few difference is present on the global results, and from one video content to another one, not always the same component achieves the best performance, even though the perceptual achromatic component (A) appears the most relevant one on two out of the four sets.

Once again, when comparing Table 2 and Table 4, the combination of the three components allows to increase significantly the performances obtained with the best single component. These results are well in accordance with more general works on human perception [39] pointing out the relative complementation of luminance and chrominance for video quality perception.

c) Sensitivity to the NN topology

As mentioned in section V, the NN architecture is defined with some meta-parameters that are related to its topology and hence influence the performances and at the same time the size of this learning machine. However, we have found quite easily many different configurations, which are reported in Table 5. They allow to vary the number of free parameters in a wide range and for which the behavior of the system is quite similar.

Table 5. Sensitivity with respect to the NN topology, results on the *LOO* global test set

Size of receptive field: *field*	Temporal delay: *delay*	Number of neurons for one receptive field: *nb_feat*	Number of neurons in the MLP hidden layer: *nb_neurons*	Number of free parameters	Mean quadratic error: *Jrmse*	Linear Correlation: *LCC*	Outlier Ratio *OR %*
25	10	5	50	5 856	0.092	0.89	18.8
12	8	12	50	12 569	0.089	0.90	17.5

20	5	20	50	31 721	0.091	0.90	17.2
20	5	20	100	53 821	0.084	0.92	15.5

The general tendency is the decreasing of the *Jrmse* cost function on the test set, it is the stochastic gradient of this objective function that is used to train the NN, while the capacity of the machine increases, meaning that over fitting has been avoided. Results presented in Table 2, Table 3, and Table 4 were obtained with the architecture corresponding to the last row of Table 5.

In all the previous experiments, the value of parameter T, which defines the size of the temporal observation sequence, see Figure 3 and Figure 5, has been set to a constant value corresponding to 5 seconds. In Table 6 and Figure 8, we report on experiments showing the influence of the length of this temporal parameter. From Table 6, it can be observed that the two smallest values ($T = 2s$, $T = 3s$) deteriorate the performances of the video quality assessment (higher error, lower correlation), while the two highest values ($T = 4s$, $T = 5s$) give comparable results, with however a slightly better behavior for $T = 5s$, which has been used in the previous experiments. A more detailed analysis shows that, according to the video content, it is either $T = 4s$ for *Cooking* or $T = 5$s for the three others that produced the best results.

Hence, we assume that $T = 5\ s$ is a reasonable upper bound, and that beyond this limit no more influence on the perceived visual quality could reasonably be awaited. Furthermore, the longer is the observation sequence, the bigger is the resulting NN architecture, consequently, it is wise not to choose a too high limit. Conversely, a lower bound will have the desirable effect of downsizing the NN architecture but at the risk of a coarse modeling of the temporal human reaction (response time and recency effect) with respect to disturbances.

Table 6. Sensitivity w.r.t. the length of the observation sequence, results on the *LOO* global test set

Length of the observation sequence *T s / # frames*	Number of free parameters	Mean quadratic error: *Jrmse*	Linear Correlation: *LCC*	Outlier Ratio *OR %*
2 s / 50	23 821	0.098	0.90	17.0
3 s/ 75	33 821	0.098	0.88	20.9
4 s / 100	43 821	0.087	0.91	17.1
5 s /125	53 821	0.084	0.92	15.5

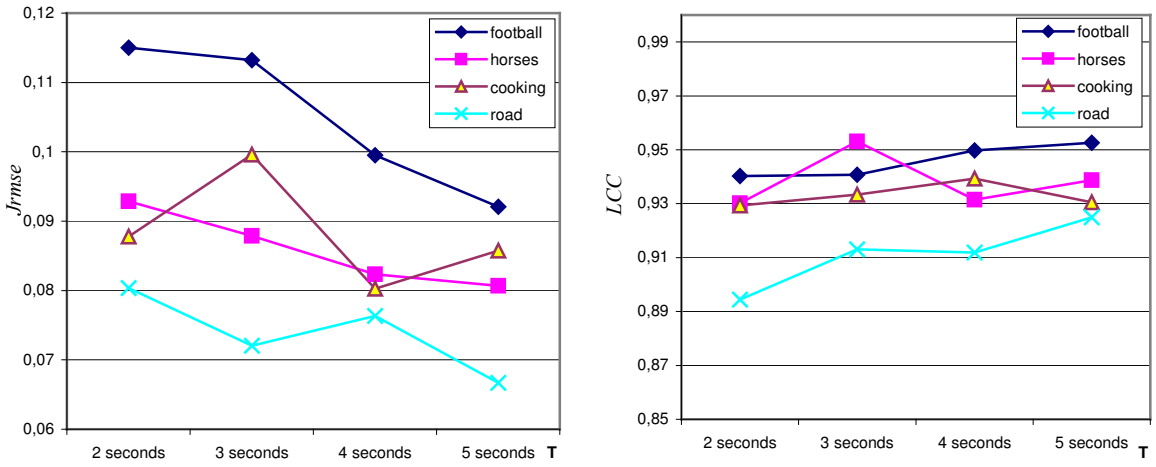


Figure 8. *Jrmse* and *LCC* w.r.t. the length of observation sequence T on the *Loo* test set.

VIII Conclusion

In this paper, we have demonstrated that TDNN can be useful to assess the perceived quality of video sequences by realizing a non-linear mapping between non subjective features extracted from the video frames and subjective scores obtained with SSCQE protocol. The proposed architecture relies on the set of convolutional neurons, which slide along the time axis sharing the same set of weights. It allows to perform not only the time integration function but also to mimic a systematic local analysis and comparison of content based features.

We have validated our approach using quite a large database that is composed of different video contents and different bit rates. Nevertheless, the main contribution compared to metrics of the literature, takes place in a way to tackle the variation of the response time of observers. This allows to the metric to perform well using usual performance measures

comparing with equivalent literature metrics. On the test set, which was independent of the learning set, a global linear correlation criteria of 0.92 (from 0.93 to 0.95 on the individual test sets) has been obtained between the output of the RR system and the subjective score provided by human observers. The outlier ratio at twice the interval of confidence on DMOS varies from about 10% to 20%, with an average value around 15%.

We have in mind to extend this system along two directions. One would be to take into account more general degradations than those due to lossy compression algorithms. Specifically, a complementary set of features sensitive to transmission errors has to be defined, and of course, for the training purpose, a new database including such kind of errors should be available. The second extension consists in replacing the spatial integration that is carried out during the feature extraction process by a learning stage that will be incorporated in the neural architecture. The same kind of approach, with convolutional neurons could be used. It leads to Space Displacement Neural Network (SDNN), which has already been used with success and combined with TDNN, for example for combining offline and online representations of handwriting [40]. Finally, we are also currently considering the evolution of this system to a full NR system.

The fields of NR and RR video quality assessment are very young, and there are many possibilities for the development of innovative metrics. We hope that the proposed combination of a TDNN, providing a statistical time-dependent model of distortions, will be useful for searchers who work in that field, specifically for those defining new features, to provide them a quite simple tool to carry on with the pooling stage.

Acknowledgment:

The authors wish to thank TDF for providing the databases used in the experiments related in this paper and Fabrice Alleau, Emilie Poisson and Marti Vilarnau for their assistance in performing the experiments described in the paper.

References:

[1] ITU-R Recommendation BT.500, "Methodology for the Subjective Assessment of the Quality of Television Pictures", http://www.itu.int/, 1999.

[2] J. Baina, G. Goudezeune, "Equipment and Strategies for Service Quality Monitoring for Digital Television Networks", *SMPTE J.*, pp. 108-632, 1999.

[3] 2. S. Wolf and M. H. Pinson, "Spatio-temporal distortion metrics for in-service quality monitoring of any digital video system," Proc. SPIE, vol. 3845, pp. 266-277, 1999.

[4] P. Le Callet, D. Barba, "Image quality assessment : from site errors to a global appreciation of quality" In proc. Picture Coding Symposium, Seoul, 2001.

[5] M. Miyahara, K. Kotani, V.R. Algazi, "Objective picture quality scale (PQS) for image coding", *IEEE Trans. Communications*, Vol. 46, No.9, pp.1215-1225, 1998.

[6] H.R. Wu, M. Yuen, "A generalize block-edge impairment metric for video coding", *IEEE Signal Processing Letters*, Vol. 4, No.11, pp.317-320, 1997.

[7] S.A. Karunasekera, N.G. Kingsbury, "A distortion measure for blocking artifacts in image based on human visual sensitivity", *IEEE Trans on Image Processing*, Vol. 4, No. 6, pp.713-724, 1995.

[8] Y. Horita, J. Ohnishi, T. Murai, "Quality evaluation model for coded JPEG2000 still pictures", *Proceedings of Picture Coding Symposium*, pp 125-128, 2001.

[9] D. S. Turaga, Y. Chen, J. Caviedes, "No reference PSNR estimation for compressed pictures", *Signal Processing: Image Commun.*, vol. 19, pp. 173–184, 2004.

[10] S. Winkler, "Issues in vision modeling for perceptual video quality assessment", *Signal Processing* 78(2), pp. 231–252, 1999.

[11] M. D'Zumra, T. J. Shen, W. Wu, H. Chen, M. Vassiliou, "Contrast gain control for color image quality", *SPIE* vol. 3299, pp 194-201, 1998.

[12] A.B. Watson, J. Hu, J.F. McGowan, "III, DVQ: A digital video quality metric based on human vision", *Journal of Electronic Imaging*, Vol. 10, No. 1, pp. 20-29, 2001

[13] M. Carnec, P. Lecallet, D. Barba, "A new method for perceptual quality assessment of compressed images with reduced reference", *SPIE* Visual Communication and Image Processing, Lugano, 2003.

[14] Z. Wang, L. Lu, A. C. Bovik, "Video quality assessment based on structural distortion measurement", *Signal Processing: Image Commun.*, vol. 19, pp. 121–132, 2004.

[15] K. T. Tan, M. Ghanbari, D. E . Pearson, "An objective measurement tool for MPEG video quality" Signal Processing, vol.. 70, pp. 279-294, 1998.

[16] M. A. Masry, S. S. Hemami, "A metric for continuous quality evaluation of compressed video with severe distortions", *Signal Processing: Image Commun.*, vol. 19, pp. 133–146, 2004.

[17] S. Yao, W. Lin, Z. Lu, E. Ong, X. Yang, "Video quality assessment using neural network based on multi-feature extraction". *Proc. SPIE Vol. 5150*, Visual Communications and Image Processing, Touradj Ebrahimi, Thomas Sikora, Eds, pp. 604-612, 2003.

[18] P. Gastaldo, S. Rovetta, R. Zunino, "Objective Quality Assessment of MPEG-2 Video Streams by Using CBP Neural Networks", *IEEE Trans. on Neural Networks*, VOL. 13, NO. 4, pp. 939-947, 2002.

[19] P. Le Callet , D. Barba"Perceptual color image quality metric using adequate error pooling for coding scheme evaluation" SPIE Human Vision and Electronic Imaging Conference,San Jose, Californie,2002.

[20] P. Le Callet , D. Barba "Robust approach for color image quality assessment", in VCIP (Visual Communications and Image Processing), Lugano, 2003.

[21] M. Carnec, P. Le Callet , D. Barba "An image quality assessment method based on perception of structural information" ICIP (International Conference on Image Processing),Barcelone,Septembre,2003.

[22] D. R. Williams, J. Krauskopf, D. W. Heeley, "Cardinal directions of color space", Vision Research, Vol. 22, pp. 1123-1131, 1982.

[23] P. Le Callet, A. Saadane, D. Barba, "Orientation selectivity of opponent-colour channels", in PERCEPTION ,vol. 28 supplement European Conference on Visual Perception, Trieste, Italy, August 22-26, 1999.

[24] R. Aldridge, D. Pearson, "A calibration method for continuous video quality (SSCQE) measurements", *Signal Processing: Image Commun.*, vol. 16, no. 3, pp. 321–332, 2000.

[25] D. Melcher, S. Wolf, "Objective Measures for Detecting Digital Tiling", *Document Number: T1A1.5/95-104*, http://www.its.bldrdoc.gov, 1995.

[26] A. A. Webster, C. T. Jones, M. H. Pinson, S. D. Voran, and S. Wolf, "An objective video quality assessment system based on human perception," Proc. SPIE, vol. 1913, pp. 15-26, 1993.

[27] T. Yamashita, M. Kameda, M. Miyahara, "An Objective Picture Quality Scale for Video Images (PQSvideo) – Definition of Distortion Factors", In Visual Communications and Image Processing 2000, *Proceedings of SPIE,* volume 4067, pp. 801-809, 2000.

[28] S. Winkler, A. Sharma, D. McNally, "Perceptual Video Quality and Blockiness Metrics for Multimedia Streaming Applications", *in Proc. 4th International Symposium on Wireless Personal Multimedia Communications,* Aalborg, Denmark, September 9-12, (invited paper), pp. 553-556, 2001.

[29] Z. Wang, A. Bovik, B. Evans, "Blind Measurement of Blocking Artefact in Images", *in Proc. ICIP*, vol. 3, Vancouver, Canada, pp. 981–984, 2000.

[30] M. Egmont-Petersena, D. de Ridder, H. Handelsc, "Image processing with neural networks—a review", *Pattern Recognition* 35, pp. 2279-2301, 2002.

[31] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-Based Learning Applied to Document Recognition", *Proc. IEEE*, vol. 86, no. 11, pp. 2278-2324, 1998.

[32] C. Garcia, M. Delakis, "Convolutional Face Finder: A Neural Architecture for Fast and Robust Face Detection", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 26, NO. 11, Nov. 2004.

[33] A. Waibel, T. Hanazawa, G. Hinton, K. Shikan, K. Lang, "Phoneme recognition using time-delay neural networks", *IEEE Trans. on Acoustics, Speech and Signal Processing*, vol. 3, pp. 328-339, 1989.

[34] M. Schenkel, I. Guyon, D. Henderson, "On-line cursive script recognition using Time Delay Neural Networks and Hidden Markov Models", *Machine Vision Application*, vol. 8, N° 4, 1995.

[35] D. Hubel, T. Wiesel, "Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex," *J. Physiology*, vol. 160, pp. 106-154, 1962.

[36] K. Fukushima, "Cognitron: A Self-Organizing Multilayered Neural Network," *Biological Cybernetics*, vol. 20, pp. 121-136, 1975.

[37] M. Mozer, "The Perception of Multiple Objects: A Connectionist Approach," *Connectionism in Perspective*, Cambridge, Mass.: MIT Press-Bradford Books, 1991.

[38] C.M. Bishop, "Neural Networks for Pattern Recognition", *Oxford University Press*. ISBN 0-19-853849-9, pages 116-161, 1995.

[39] A.B. Watson, J.A. Soloman, "A model of visual contrast gain control and pattern masking", *Journal of the Optical Society of America* A, Vol. 14, No. 9, pp. 2379-2391, 1997.

[40] E. Poisson, C. Viard-Gaudin, P.M. Lallican, "Multi-modular architecture based on convolutional neural networks for online handwritten character recognition", *In proc. ICONIP*, Singapore, 2002.